Analysis of MALDI-TOF Serum Profiles for Biomarker Selection and Sample Classification

H. W. Ressom[*1], R. S. Varghese[1], E. Orvisky[1], S. K. Drake[2], G. L. Hortin[2],
M. Abdel-Hamid[3], C. A. Loffredo[1], and R. Goldman[1]

[1]Lombardi Comprehensive Cancer Center, Georgetown University, Washington, DC
[2]Clinical Chemistry Service, Department of Laboratory Medicine, NIH, Bethesda, MD
[3]Viral Hepatitis Research Laboratory, NHTMRI, Cairo, Egypt
[*]Corresponding author: hwr@georgetown.edu

***Abstract*- Mass spectrometric profiles of peptides and proteins obtained by current technologies are characterized by complex spectra, high dimensionality, and substantial noise. These characteristics generate challenges in discovery of proteins and protein-profiles that distinguish disease states, e.g. cancer patients from healthy individuals. A challenging aspect of biomarker discovery in serum is the interference of abundant proteins with identification of disease-related proteins and peptides. We present data processing methods and computational intelligence that combines support vector machines (SVM) with particle swarm optimization (PSO) for biomarker selection from MALDI-TOF spectra of enriched serum. SVM classifiers were built for various combinations of m/z windows guided by the PSO algorithm. The method identified mass points that achieved high classification accuracy in distinguishing cancer patients from non-cancer controls. Based on their frequency of occurrence in multiple runs, six m/z windows were selected as candidate biomarkers. These biomarkers yielded 100% sensitivity and 91% specificity in distinguishing liver cancer patients from healthy individuals in an independent dataset.**

I. Introduction

Mass spectrometric serum profiling was optimized for high-throughput comparison of complex samples that allows discovery of biomarkers of diseases such as cancer [1]. Independent analysis of the results pointed out the importance of avoiding bias and the need for independent validation of results [2-4]. Improved study design and technology in second-generation studies continue to indentify biomarker-candidates for variety of cancers [5-7]. This paper adds data preprocessing and feature selection methods to a growing number of improved tools for matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometric identification of biomarkers in enriched serum.

Mass spectra represent a complex signal consisting of electronic noise, chemical noise due to contaminants and matrix, and protein and metabolic signatures [8]. They also have a varying baseline caused, besides others, by matrix-associated chemical noise or by ion overload. The latter refers to the high excess of ions derived from the matrix that can overload the detector [9]. This elevates the baseline from its ideal zero horizontal line. Previous quality-control experiments have suggested several measurement properties of current mass spectrometry technologies that must be accounted for in the analysis [10]. These properties include high dimensionality of the spectra, high coefficients of variation, and mass shift (measurement error). Thus, it is important to apply preprocessing methods that enable the recognition of spectral quality prior to using the spectra for biomarker discovery and sample classification.

Data preprocessing methods such as smoothing, baseline correction, normalization, peak detection, and peak alignment improve the performance of mass spectrometric data analysis methods for biomarker discovery [9, 11]. The reason for this includes the substantial amount of noise and systematic variations between spectra caused by sample degradation over time, ionization suppression, and other parameters reviewed previously by [4, 12]. The data preprocessing methods are typically available in all software for operation of a mass spectrometer. The use of spectral comparisons for biomarker identification requires, however, optimization of these methods and a completely transparent data manipulation. Several groups proposed recently improved tools for data preprocessing and biomarker discovery as summarized briefly below.

By smoothing the raw spectra, we can reduce the effect of some mass-per-charge (m/z) values that appear as peaks but may not be or are very hard to verify by independent experiments. Many smoothing algorithms are available to denoise raw signals including the well-known Savitzky-Golay filter that removes additive white noise [13] and wavelets [14].

Baseline correction is important for minimization of background noise; drifting baseline introduces serious distortion of ion intensities without adequate correction. Several methods have been proposed for baseline subtraction. For example, Fung and Enderwick [15] employed a varying-width segemented convex hull algorithm to subtract the baseline. Baggerly *et al.* [16] fitted a local median or local mean in a fixed window on the time scale. They also considered subtracting a “semimonotonic” baseline. Coombes *et al.* [14] estimated baseline by fitting a monotone local minimum curve to smoothed spectra.

Normalization reduces variation in signal intensity between spectra. A commonly used normalization method for mass spectrometric data is rescaling each spectrum by its total ion current, i.e., the area under the curve (AUC) [11, 15]. Other common choices for the rescaling coefficient include the

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE **2005**	2. REPORT TYPE	3. DATES COVERED **00-00-2005 to 00-00-2005**

4. TITLE AND SUBTITLE **Analysis of MALDI-TOF Serum Profiles for Biomarker Selection and Sample Classification**	5a. CONTRACT NUMBER 5b. GRANT NUMBER 5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER 5e. TASK NUMBER 5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Georgetown University,Lombardi Comprehensive Cancer Center,3970 Reservoir Road Northwest,Washington,DC,20057**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S) 11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**7**	

spectrum median or mean. Alternatively, choosing the average AUC over all spectra as the rescaling coefficient can do a global normalization. A global optimization assumes that the sample intensities are all related by a constant factor. That means that the data distribution should not differ substantially from one spectrum to another.

Peak detection deals with the selection of m/z values which display a reasonable intensity compared to those that appear as noise. Coombes *et al.* [14] applied a simple peak finding (SPF) algorithm that provides the locations of potential peaks and their associated left-hand and right-hand bases. They estimated signal-to-noise ratio (S/N) using wavelets for improved peak detection. Also, they introduced a method for coalescing neighboring peaks.

Assuming appropriate mass spectral data preprocessing methods are used, biomarker selection can be addressed using various computational methods. One of the commonly used approaches is to apply statistical analyses that recognize differentially expressed m/z values between cases and controls with multiple subjects. For example, one can apply a two-sample t-test method to compare the protein intensities at each m/z value in cases and controls. Zhu *et al.* [17] proposed a statistical algorithm that can select a subset of *k* biomarkers from the marker list that could best discriminate between the groups in a training dataset via the best *k*-subset discriminant method with high sensitivity and specificity.

Computational intelligence has also been applied for biomarker discovery. For example, Petricoin *et al.* [1] used a combination of genetic algorithm (GA) and self-organizing clustering (GA-SOC) for variable selection. The GA-SOC, which is implemented in ProteomeQuest software, starts with hundreds of random choices of small sets of exact m/z values selected from the SELDI-TOF mass spectra. Each candidate subset contains 5 to 20 of the potential m/z values that define the spectra. The m/z values within the highest rated sets are reshuffled to form new subset candidates. The candidates are rated iteratively until the set that fully discriminates the preliminary set emerges.

Koopmann *et al.* [18] applied successfully support vector machines (SVMs) in a modified form to proteomic profiling. Li *et al.* (2002) introduced unified maximum separability analysis (UMSA) algorithm, which incorporates data distribution information into structural risk minimization learning algorithm. UMSA is applied to identify a direction along which two classes of data are best separated. This direction is represented as a linear combination of the original variables. The weight assigned to each variable in this combination measures the contribution of the variable toward the separation of the two classes of data. They analyzed protein profiles of serum samples from patient with or without breast cancer. They reported that UMSA enabled the identification of three discriminatory biomarkers that achieved 93% sensitivity and 91% specificity in detecting breast cancer patients from the non-cancer controls.

In our previous work [19, 20], we proposed a novel computational method known as PSO-SVM that combines SVMs and particle swarm optimization (PSO) for optimal selection of m/z values from high resolution surface enhanced laser desorption ionization-quadrupole time-of-flight (SELDI-QqTOF) spectra. In [20], we performed binning, normalization, baseline correction, peak identification, and peak alignment on SELDI-QqTOF spectra. The peak alignment method defines windows of m/z values that have variable width. The PSO-SVM algorithm is then applied to select the optimal m/z windows. We ran the algorithm multiple times and selected five m/z windows based on their frequency of occurrence. An SVM classifier that employs these five m/z windows as its inputs yielded 92% sensitivity and 90% specificity in distinguishing hepatocellular carcinoma (HCC) patients from matched controls.

In this paper, the serum samples were enriched by denaturing ultrafiltration and desalting [21] on C8 magnetic beads (MB) [22]. The procedure disrupts protein-protein interactions and allows an efficient recovery of a low molecular weight (LMW) serum fraction starting with 25 µl of serum. The enrichment offers more peaks than unenriched SELDI-QqTOF or unenriched MALDI-TOF spectra [23]. This paper presents our analysis of MALDI-TOF spectra of enriched serum for biomarker discovery and sample classification.

II. Methods and Results

A. Mass Spectral Data

The incidence of HCC in the United States increases. Very high rates of HCC incidence are observed in Egypt where an epidemic of viral infections presents a serious health problem. The management of the disease would benefit from identification of biomarkers related to this disease. Serum samples of HCC cases and controls were obtained from 2000 to 2002 in collaboration with the National Cancer Institute of Cairo University, Egypt. Controls were recruited among patients from the orthopedic fracture clinic at the Kasr El-Aini Hospital, Cairo, Egypt and were frequency-matched to cancer cases by gender, rural versus urban birthplace, and age [24]. Blood samples were collected by trained phlebotomist each day around 10am and processed within a few hours according to a standard protocol. Aliquots of sera for mass spectrometric analysis were frozen at -80^{o}C immediately after collection till analysis; all measurements were performed on samples of second-time thawed serum.

Eluted peptides in the enriched serum samples were mixed with a matrix solution (3 mg/ml α-cyano-4-hydroxycinaminic acid in 50% actonitrile with 0.1% trifluoracetic acid), spotted onto AnchorChip target (Bruker Daltonics, Billerica, MA), and analyzed using an Ultraflex MALDI TOF/TOF mass spectrometer (Bruker Daltonics, Billerica, MA). Each spectrum was detected in linear positive mode and was externally calibrated using a standard mixture of peptides. Denaturing ultrafiltration enriches LMW fraction of serum and plasma (Fig. 1). Removal of proteins greater than 50 kDa including albumin appears complete based on Coomassie staining; proteins smaller than 50kDa are also removed as shown by the SDS-PAGE in Fig. 1 (left). Fig. 1 (right) depicts the spectrum found after desalting (top spectrum) and after denaturing ultrafiltration (bottom spectrum). The enrichment improved the quality of the spectrum in the LMW region and provided over 300 peaks. Evidently, the enrichment (bottom spectrum) offers more peaks than an unenriched spectrum (top spectrum).

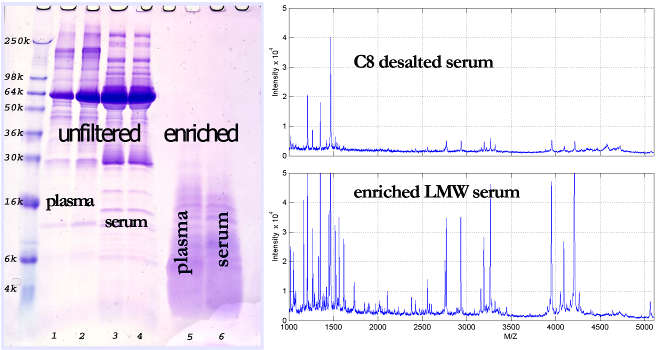


Fig. 1. *Left:* SDS-PAGE analysis of human plasma and serum. *Lane 1 and 2*, unfiltered plasma, *Lanes 3 and 4*, unfiltered serum, *lane 5*, enriched LMW plasma and *lane 6*, enriched LMW serum. 10 µg of total protein was applied per lane and visualized by Coomassie staining. *Right*: MALDI-TOF spectrum after desalting using C8 magnetic beads (top spectrum) and after denaturing ultrafiltration. (bottom spectrum).

B. Reproducibility

Our study used 62 replicate spectra to examine the reproducibility of MALDI-TOF mass spectrometry. Each spectrum consisted of ~136,000 m/z values with the corresponding ion intensities. The dimension of these high-resolution spectra was reduced to 23,846 m/z values using the binning procedure that divides the m/z axis into intervals of desired length over the mass range 0.9 to 10 kDa. A bin size of 100 parts per million (ppm) was found adequate. The mean of the intensities within each interval was used as the protein expression variable in each bin. Each intensity value was transformed by computing the base-two logarithm and found the mean log intensity value and standard deviation.

The CV of the log-transformed intensity values in the 62 reference spectra ranged between 4.1% and 22.9% with mean value of 10.5%. A heat map for 62 replicate spectra (Fig. 2) and spectra for three independently prepared samples of enriched LMW fraction of serum (Fig. 3) illustrate the reproducibility of MALDI-TOF MS.

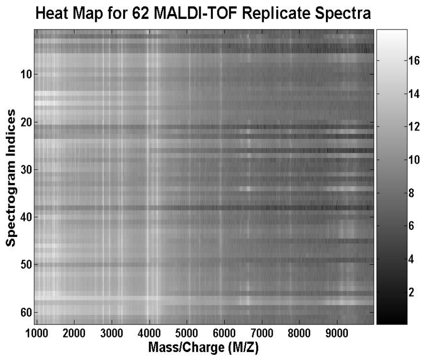


Fig 2. Heat map for 62 MALDI-TOF replicate spectra.

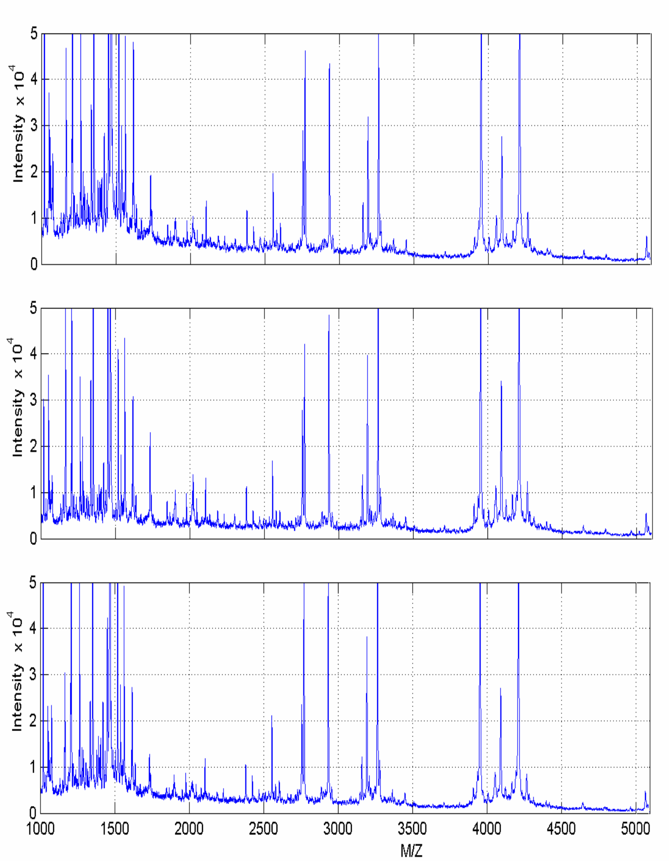


Fig 3. MALDI-TOF spectra, three independently prepared samples of enriched LMW fraction of serum.

C. Data Preprocessing

We applied various methods to preprocess the raw MALDI-TOF mass spectra. We began our analysis with outlier screening where spectra whose data distribution substantially deviated from others were removed. 14 of the 164 MALDI-TOF spectra were excluded, leaving 150 (78 cases and 72 controls) serum mass spectral profiles for further analysis. These outliers were singled out based on their deviation from the median ion current, median record count (number of mass points), and their alignment with pre-selected landmarks.

Each spectrum was first binned with a bin size of 100 ppm, which reduced the dimension of the spectra from about 136,000 m/z values to 23,846 bins over the mass range 0.9 to 10 kDa. Figure 4a and 4b depict a typical raw spectrum of a healthy individual and the corresponding binned spectrum, respectively. On the horizontal axis are m/z values or bins and on the vertical axis are intensity measurements that indicate the relative ion abundance. As shown in the figures, the binning algorithm has removed some high frequency noise, thus smoothing the spectrum. Also, binning improves the alignment of multiple spectra (not shown).

The baseline of each binned spectrum was estimated by obtaining the minimum value within a shifting window size of 50 bins. Spline approximation was used to regress the varying baseline. The regressed baseline was subtracted from the spectrum yielding a baseline corrected spectrum. Spline regression estimates different linear slopes for different ranges of the m/z values. Eilers and Marx [25] applied the method for baseline correction of 2-D gel electrophoresis images. Furthermore, each spectrum was normalized by dividing it by its total ion current. In addition, the spectra were scaled to have an overall maximum intensity of 100. Fig. 4c shows the binned, normalized, and baseline corrected spectrum.

For peak detection, a bin is identified as a peak if the sign of the intensity's slope changes from positive to negative. Peaks with intensity below a pre-defined threshold-line were considered as noise and were discarded. We selected m/z values with reasonable intensity levels and discarded those that appeared as noise. Following outlier screening, binning, baseline correction, normalization, and peak detection, the 78 HCC case and 72 control spectra were split into 100 training spectra (50 HCC and 50 normal) and 50 testing spectra (28 HCC and 22 normal). The testing spectra were scaled based on the parameters used for scaling the training spectra.

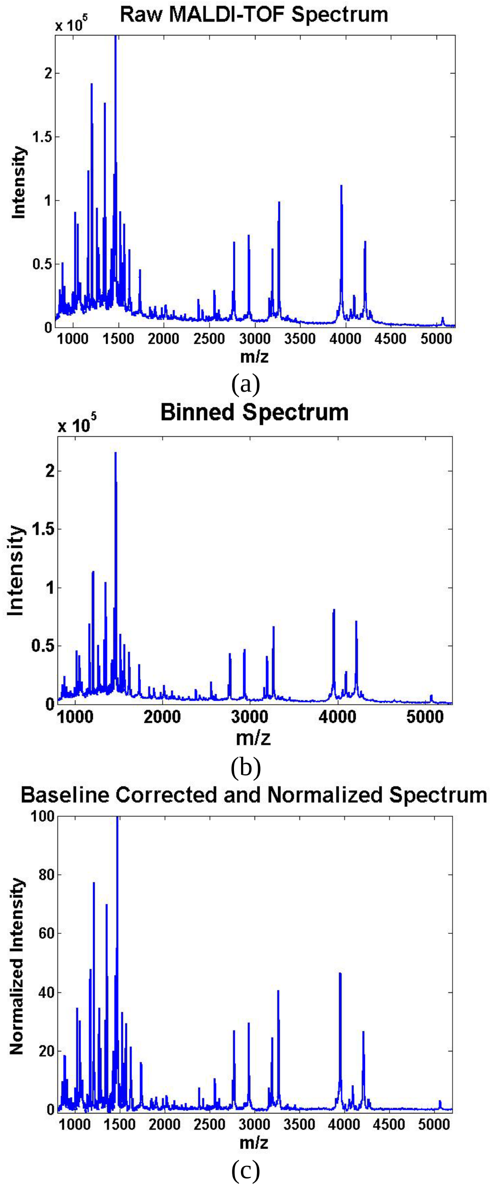
To account for variation in the m/z location (drifts) in different spectra, two peaks were coalesced if they differed in location by at most 7 bins or at most 0.03% relative mass. This method was based on the ideas of Coombes *et al.* [14] who used this method for SELDI-TOF spectra, where they combined peaks if they fall within 7 clock ticks and differ by at most 0.3% relative mass. We applied this method on training dataset only and found 264 windows. Fig. 5 shows m/z windows found between 1730 and 1870 Da. For each spectrum, the maximum intensity within each window was found, yielding a 264 x 100 training data matrix. The same windows were used to quantify the peaks in the testing spectra, which resulted in a 264 x 50 testing data matrix.

D. PSO-SVM

The PSO-SVM algorithm can be used to identify optimal m/z windows from preprocessed mass spectra. While PSO selects subsets of predefined m/z windows as potential solutions, SVM classifiers are built for each potential solution generated by PSO. The prediction capability of the resulting SVM classifier on a validation dataset is used as a performance function for the PSO algorithm. Since SVMs provide good generalization capability in classification tasks and can be designed in a computationally efficient manner, they are an ideal candidate for use as a performance function.

The training dataset is used to select m/z windows and build an SVM classifier. The validity of each classifier trained with the selected features is evaluated using the prediction accuracy of the SVM classifier in distinguishing cancer patients from non-cancer controls. SVM classifiers are built for various combinations of features until the performance of the SVM classifier converges or a pre-specified maximum iteration number is reached.

Estimates of prediction accuracy are calculated by using the *k*-fold cross-validation and bootstrapping methods. In *k*-fold cross-validation, we divide the training dataset into *k* subsets of (approximately) equal size. We train the SVM classifier *k* times, each time leaving out one of the subsets from training, but using only the omitted subset to compute the prediction accuracy.



(a)

(b)

(c)

Fig. 4. MALDI-TOF spectrum of a standard serum sample processed by smoothing, baseline correction, and normalization. (a) raw ; (b) binned; and (c) binned, normalized, and baseline corrected spectrum.

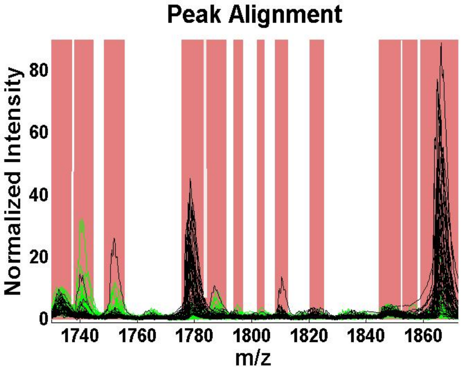


Fig. 5. Control spectra (black), case spectra (light), windows in the m/z range from 1.73 to 1.87 kDa.

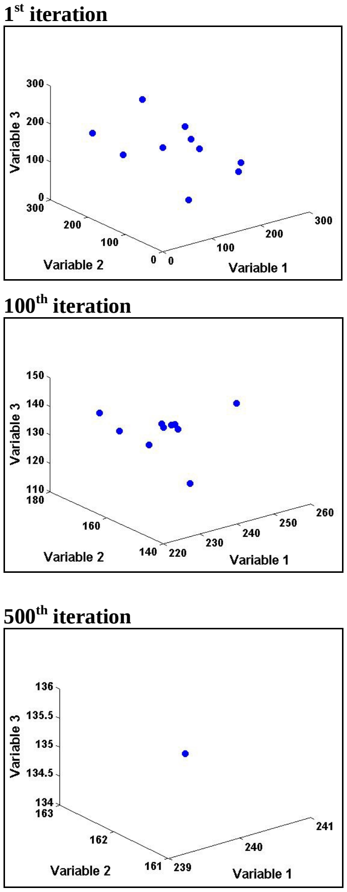
In bootstrapping, instead of analyzing pre-specified subsets of the training dataset, we repeatedly select subsamples of the data. Each subsample is a random sample with replacement from the full training dataset.

The PSO-SVM algorithm is used to identify the optimal m/z windows from a list of *L* potential m/z windows. The algorithm creates *N* vectors (particles), each consisting of *n* m/z windows that are randomly selected from *L* m/z windows. The algorithm evaluates the performance of each particle in distinguishing cancer cases from controls. This is carried out by building an SVM classifier for each particle and evaluating the performance of the classifier via the *k*-fold cross-validation or bootstrapping methods. The algorithm uses the most-fit particles to contribute to the next generation of *N* candidate particles. Thus, on the average, each successive population of candidate particles fits better than its predecessor. This process continues until the performance of the SVM classifier converges.

The algorithm repeats the above steps multiple times and provides a list of selected m/z windows along with their frequency of occurrence. A frequency plot is used to estimate the optimal number of m/z windows. The frequency plot presents the number of occurrences versus the m/z windows sorted in the order of decreasing frequency. We considered as candidate biomarkers all m/z windows starting from the first until the frequency curve becomes flat (i.e. the change in frequency becomes low). These m/z windows are evaluated via testing dataset (i.e., independent dataset that was used neither for training nor for variable selection) to determine the generalization capability of the SVM classifier.

We present as an example a single run to demonstrate how the PSO-SVM algorithm selects three markers (n=3) out of 264 m/z windows (*L*=264) using 100 MALDI-TOF spectra. The number of particles in this example is 10 (*N*=10). Note that the algorithm searches in a continuous search space but the numbers are rounded to the nearest integer. The elements of a particle represent the variable set suggested by the particle. Each particle is used to build an SVM classifier. In this example, the performance of the SVM classifier is evaluated through the bootstrapping method that randomly splits the spectra (80% for building an SVM classifier and the remaining 20% for validation). This is repeated 500 times with resubstitution and the average prediction accuracy on the validation set is computed.

Fig. 6 shows the variable sets selected and their prediction accuracy on the validation set at the 1st, 100th, and 500th iterations, respectively. The left panel depicts the location of the particles in a three-dimensional space. The tables in the right panel show the corresponding coordinates sorted in decreasing order of their prediction accuracy (only the top three and the bottom two variable sets among the 10 variable sets are presented). As shown in the figure, the particles converged to one location (240, 162, 135) after 500 iterations improving the prediction accuracy from 77% to 91%. This location corresponds to m/z windows 4644.9-4651.4, 2528.7-2535.5, and 1863.4-1871.3.



1st iteration

Selected Variable Sets			Accu. %
239	213	90	77
156	257	230	73
75	25	139	72
.	.	.	.
.	.	.	.
99	234	115	60
172	224	112	54

100th iteration

Selected Variable Sets			Accu. %
240	162	135	91
239	162	135	90
237	162	135	85
.	.	.	.
.	.	.	.
227	144	146	63
228	173	138	58

500th iteration

Selected Variable Sets			Accu. %
240	162	135	91
240	162	135	91
240	162	135	91
.	.	.	.
.	.	.	.
240	162	135	91
240	162	135	91

Fig. 6. Variable sets selected by the PSO-SVM algorithm and their prediction accuracy at the 1st, 100th, and 500th iterations. The figures in the left panel show the location of the particles in the three-dimensional space. Each table in the right panel shows the top three and the bottom two variable sets among the 10 variable sets (particles) used by PSO, sorted in decreasing order of prediction accuracy.

E. Biomarker Selection

The purpose of this analysis is to identify optimal m/z windows or candidate biomarkers from the preprocessed mass spectral data. While peak detection deals with the selection of mass points with reasonable intensity and S/N ratio, the aim of biomarker selection is to identify mass points that can be used to distinguish between cancer patients and healthy individuals.

We used the PSO-SVM algorithm to select candidate biomarkers from the 264 peak-containing m/z windows. In this study, we arbitrarily targeted selection of five m/z windows. The algorithm began with 100 particles where each particle consisted of 5 randomly selected m/z values from the 264 windows (i.e., *n* = 5, *N* = 100, and *L* = 264). A linear SVM classifier was built for each particle via the training dataset. The prediction power of each particle (five m/z windows) was evaluated by measuring the performance of the SVM classifier in distinguishing the two classes through the *k*-fold cross validation and bootstrapping methods. We used *k*=10 for this study. The most-fit particles contributed to the next generation of 100 candidate particles. This process continued until the performance of the SVM classifier converged or a pre-specified number of iterations was reached. The algorithm was repeated 600 times, of which about half of the runs used the 10-fold cross-validation method and the other half used the bootstrapping method. Fig. 7 depicts the percentage of occurrence of m/z windows selected by the PSO-SVM. Note that the m/z windows are sorted in decreasing order of frequency and only the first 50 m/z windows are shown in the figure. Fig. 7 suggests that the first seven m/z windows are more frequently selected. Our TOF/TOF sequencing indicated that the first and the seventh m/z windows share the same sequence except for one amino acid. Thus, our subsequent analysis considered only the first six m/z windows. These six m/z windows yielded 100% sensitivity and 91% specificity in distinguishing liver cancer patients from healthy individuals in the testing dataset. Fig. 8 shows the box plot for the six m/z windows identified by the PSO-SVM algorithm. As shown in the figure, each of the six m/z windows is statistically significant candidate biomarkers.

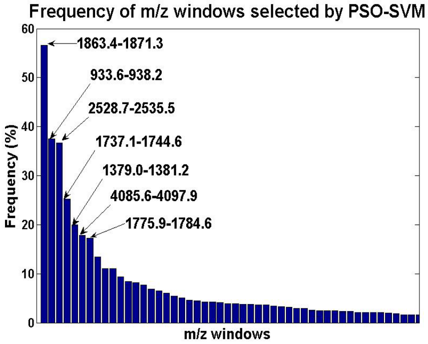


Fig. 7. Frequency of occurrence of m/z windows in 600 PSO-SVM runs for preprocessed spectra sorted in decreasing order of frequency (only the first 50 m/z windows are shown).

To examine the effect of data preprocessing on biomarker selection and sample classification, we performed biomarker selection using spectra that were binned and normalized, but not baseline corrected. 292 m/z windows were found from these spectra using our peak detection and alignment methods described before. The increase in the number of m/z windows is attributed to features that were not baseline corrected. The PSO-SVM algorithm was run 200 times with 100 particles to select 5 m/z windows out of 292 (i.e. n = 5, N = 100, and L = 292). The resulting frequency plot (Fig. 9) provided 5 biomarkers, of which the top 3 were very close to those found in the previous experiment. These 5 candidate biomarkers yielded 89% sensitivity and 86% specificity. This is significantly less than the prediction performance obtained when baseline correction was used in data preprocessing. To perform a fair comparison with the previous experiment, we tested the first six m/z windows from Fig. 9. However, the addition of the sixth m/z window did not improve the prediction accuracy. This shows that baseline correction has an impact in selecting biomarkers that provide improved sample classification.

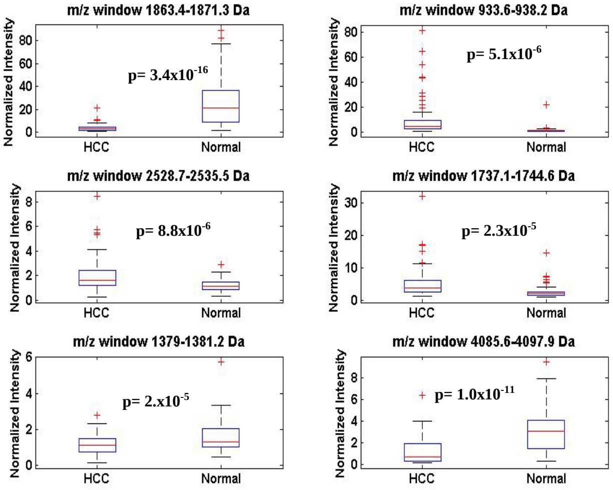


Fig. 8. Boxplots for the six m/z windows identified by the PSO-SVM. The boxplots show the distribution of each m/z window for HCC cases and normal using in both training and testing datasets combined.

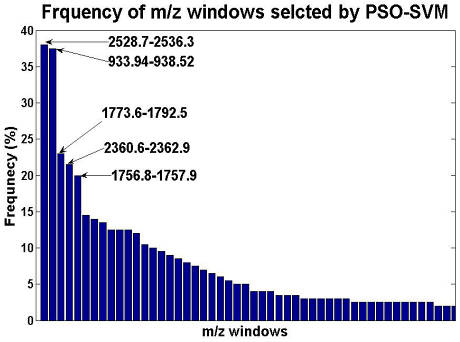


Fig. 9. Frequency of occurrence of m/z windows in 200 PSO-SVM runs for non-baseline corrected spectra, sorted in decreasing order of frequency (only the first 50 m/z windows are shown).

F. *Sample Classification*

We applied three classification algorithms, k-nearest neighbor (KNN), linear discriminant analysis (LDA), and SVMs to build classifiers. For comparison, we used three sets of features as inputs to the classifiers: all m/z bins, all m/z windows, and the six m/z windows selected by the PSO-SVM algorithm. Table 1 shows the sensitivity and specificity of the three classifiers in distinguishing HCC patients from healthy individuals in the testing dataset. Over all, the classifiers that used the six m/z windows performed better than those that used all m/z bins and m/z windows.

TABLE 1
PREDICTION ACCURACY OF THREE CLASSIFIERS ON THE TESTING DATASET.

Classification Methods	23,846 m/z bins		264 m/z windows		6 m/z windows	
	Sen.	Spec.	Sen.	Spec.	Sen.	Spec.
KNN (K=3)	96	77	96	73	93	91
LDA	89	91	89	95	98	92
SVM	93	91	93	86	100	91

III. Conclusions

This paper presents computational methods for preprocessing of mass spectral data, biomarker selection, and sample classification. Together, PSO and SVM are applied to identify candidate biomarkers from preprocessed MALDI-TOF spectra of enriched serum. The biomarkers distinguish cancer patients from non-cancer controls with high sensitivity and specificity. The PSO is used here to select a parsimonious subset from a large set of features. Since the particles contain discrete information only, we are currently investigating discrete methods such as ant colony optimization.

Acknowledgment

This work was supported in part by U.S. Army Medical Research and Material Command, Prostate Cancer Research Program grant DAMD17-02-1-0057 and American Cancer Society grant CRTG-02-245-01-CCE awarded to RG. We thank J. Jelinek for the programming of the binning algorithm. We thank Drs. F. Seillier-Moiseiwitsch and A. Wang for their helpful suggestions and discussions.

References

[1] E. F. Petricoin, A. M. Ardekani, B. A. Hitt, P. J. Levine, V. A. Fusaro, S. M. Steinberg, G. B. Mills, C. Simone, D. A. Fishman, E. C. Kohn, and L. A. Liotta, "Use of proteomic patterns in serum to identify ovarian cancer," *Lancet*, vol. 359, pp. 572-7, 2002.

[2] K. A. Baggerly, J. S. Morris, and K. R. Coombes, "Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments," *Bioinformatics*, vol. 20, pp. 777-85, 2004.

[3] E. P. Diamandis, "Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations," *Mol Cell Proteomics*, vol. 3, pp. 367-78, 2004.

[4] D. F. Ransohoff, "Bias as a threat to the validity of cancer molecular-marker research," *Nat Rev Cancer*, vol. 5, pp. 142-9, 2005.

[5] T. P. Conrads, V. A. Fusaro, S. Ross, D. Johann, V. Rajapakse, B. A. Hitt, S. M. Steinberg, E. C. Kohn, D. A. Fishman, G. Whitely, J. C. Barrett, L. A. Liotta, E. F. Petricoin, 3rd, and T. D. Veenstra, "High-resolution serum proteomic features for ovarian cancer detection," *Endocr Relat Cancer*, vol. 11, pp. 163-78, 2004.

[6] V. Paradis, F. Degos, D. Dargere, N. Pham, J. Belghiti, C. Degott, J. L. Janeau, A. Bezeaud, D. Delforge, M. Cubizolles, I. Laurendeau, and P. Bedossa, "Identification of a new marker of hepatocellular carcinoma by serum protein profiling of patients with chronic liver diseases," *Hepatology*, vol. 41, pp. 40-7, 2005.

[7] Z. Zhang, R. C. Bast, Jr., Y. Yu, J. Li, L. J. Sokoll, A. J. Rai, J. M. Rosenzweig, B. Cameron, Y. Y. Wang, X. Y. Meng, A. Berchuck, C. Van Haaften-Day, N. F. Hacker, H. W. de Bruijn, A. G. van der Zee, I. J. Jacobs, E. T. Fung, and D. W. Chan, "Three biomarkers identified from serum proteomic analysis for the detection of early stage ovarian cancer," *Cancer Res*, vol. 64, pp. 5882-90, 2004.

[8] E. F. Petricoin, 3rd, D. K. Ornstein, C. P. Paweletz, A. Ardekani, P. S. Hackett, B. A. Hitt, A. Velassco, C. Trucco, L. Wiegand, K. Wood, C. B. Simone, P. J. Levine, W. M. Linehan, M. R. Emmert-Buck, S. M. Steinberg, E. C. Kohn, and L. A. Liotta, "Serum proteomic patterns for detection of prostate cancer," *J Natl Cancer Inst*, vol. 94, pp. 1576-8, 2002.

[9] D. I. Malyarenko, W. E. Cooke, B. L. Adam, G. Malik, H. Chen, E. R. Tracy, M. W. Trosset, M. Sasinowski, O. J. Semmes, and D. M. Manos, "Enhancement of sensitivity and resolution of surface-enhanced laser desorption/ionization time-of-flight mass spectrometric records for serum peptides using time-series analysis techniques," *Clin Chem*, vol. 51, pp. 65-74, 2005.

[10] Y. Yasui, M. Pepe, M. L. Thompson, B. L. Adam, G. L. Wright, Jr., Y. Qu, J. D. Potter, M. Winget, M. Thornquist, and Z. Feng, "A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection," *Biostatistics*, vol. 4, pp. 449-63, 2003.

[11] A. C. Sauve, T. P. Speed, and "Normalization, baseline correction and alignment of high-throughput mass spectrometry data " *Proceedings of the Genomic Signal Processing and Statistics workshop, Baltimore, MD, USA.*, May 26-27, 2004.

[12] O. J. Semmes, Z. Feng, B. L. Adam, L. L. Banez, W. L. Bigbee, D. Campos, L. H. Cazares, D. W. Chan, W. E. Grizzle, E. Izbicka, J. Kagan, G. Malik, D. McLerran, J. W. Moul, A. Partin, P. Prasanna, J. Rosenzweig, L. J. Sokoll, S. Srivastava, S. Srivastava, I. Thompson, M. J. Welsh, N. White, M. Winget, Y. Yasui, Z. Zhang, and L. Zhu, "Evaluation of serum protein profiling by surface-enhanced laser desorption/ionization time-of-flight mass spectrometry for the detection of prostate cancer: I. Assessment of platform reproducibility," *Clin Chem*, vol. 51, pp. 102-12, 2005.

[13] L. Pusztai, B. W. Gregory, K. A. Baggerly, B. Peng, J. Koomen, H. M. Kuerer, F. J. Esteva, W. F. Symmans, P. Wagner, G. N. Hortobagyi, C. Laronga, O. J. Semmes, G. L. Wright, Jr., R. R. Drake, and A. Vlahou, "Pharmacoproteomic analysis of prechemotherapy and postchemotherapy plasma samples from patients receiving neoadjuvant or adjuvant chemotherapy for breast carcinoma," *Cancer*, vol. 100, pp. 1814-22, 2004.

[14] K. R. Coombes, S. Tsavachidis, J. S. Morris, K. A. Baggerly, M. C. Hung, and H. M. Kuerer, "Improved peak detection and quantification of mass spec-trometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform," The University of Texas M.D. Anderson Cancer Center, Technical Report UTMDABTR-001-04, 2004.

[15] E. T. Fung and C. Enderwick, "ProteinChip clinical proteomics: computational challenges and solutions," *Biotechniques*, vol. Suppl, pp. 34-8, 40-1, 2002.

[16] K. A. Baggerly, J. S. Morris, J. Wang, D. Gold, L. C. Xiao, and K. R. Coombes, "A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization-time of flight proteomics spectra from serum samples," *Proteomics*, vol. 3, pp. 1667-72, 2003.

[17] W. Zhu, X. Wang, Y. Ma, M. Rao, J. Glimm, and J. S. Kovach, "Detection of cancer-specific markers amid massive mass spectral data," *Proc Natl Acad Sci U S A*, vol. 100, pp. 14666-71, 2003.

[18] J. Koopmann, Z. Zhang, N. White, J. Rosenzweig, N. Fedarko, S. Jagannath, M. I. Canto, C. J. Yeo, D. W. Chan, and M. Goggins, "Serum diagnosis of pancreatic adenocarcinoma using surface-enhanced laser desorption and ionization mass spectrometry," *Clin Cancer Res*, vol. 10, pp. 860-8, 2004.

[19] H. Ressom, R. S. Varghese, D. Saha, E. Orvisky, L. Goldman, E. F. Petricoin, T. P. Conrads, T. D. Veenstra, M. Abdel-Hamid, C. A. Loffredo, and R. Goldman, "Particle swarm optimization for analysis of mass spectral serum profiles," *Proceedings of Genetic and Evolutionary Computation Conference (GECCO)*, vol. 1 pp. 431-438, 2005.

[20] H. W. Ressom, R. S. Varghese, M. Abdel-Hamid, S. Abdel-Latif Eissa, D. Saha, L. Goldman, E. F. Petricoin, T. P. Conrads, T. D. Veenstra, C. A. Loffredo, and R. Goldman, "Analysis of mass spectral serum profiles for biomarker selection," *Bioinformatics*, *in press*, 2005.

[21] R. S. Tirumalai, K. C. Chan, D. A. Prieto, H. J. Issaq, T. P. Conrads, and T. D. Veenstra, "Characterization of the low molecular weight human serum proteome," *Mol Cell Proteomics*, vol. 2, pp. 1096-103, 2003.

[22] X. Zhang, S. M. Leung, C. R. Morris, and M. K. Shigenaga, "Evaluation of a novel, integrated approach using functionalized magnetic beads, bench-top MALDI-TOF-MS with prestructured sample supports, and pattern recognition software for profiling potential biomarkers in human plasma," *J Biomol Tech*, vol. 15, pp. 167-75, 2004.

[23] E. Orvisky, S. K. Drake, B. M. Martin, M. Abdel-Hamid, H. W. Ressom, R. S. Varghese, D. Saha, G. L. Hortin, C. A. Loffredo, and R. Goldman, "Enrichment of low molecular weight fraction of serum for mass spectrometric analysis of peptides associated with hepatocellular carcinoma," *Submitted to Proteomics*, 2005.

[24] S. Ezzat, M. Abdel-Hamid, S. Abdel-Latif Eissa, N. Mokhtar, N. A. Labib, L. El-Ghorory, N. N. Mikhail, A. Abdel-Hamid, T. Hifnawy, G. T. Strickland, and C. A. Loffredo, "Associations of pesticides, HCV, HBV, and hepatocellular carcinoma in Egypt," *Int J Hygiene Env Health*, *in press*, 2005.

[25] P. H. C. Eilers and B. D. Marx, "Flexible smoothing with B-splines and penalties," *Statist. Sci.*, vol. 11(2), pp. 89–121, 1996.